Intro to R for Epidemiologists

Lab 1 (1/15/15)

Many of these questions go beyond the information provided in the lecture. Therefore, you may need to use R help files and the internet to search for answers. Feel free to ask questions of the instructor, the TAs, or your classmates, but try to work through as much as you can independently.

For the lab, you are expected to create an R script (.R file in the R editor) with your code corresponding to each question. Begin each question with a commented line of code indicating the question. As an example:

```r
# Jenna Krall

# Question 1.
head(iris)
```

Part 1. Characteristics of irises

1. The dataset `iris` in R comes with the preloaded `datasets` R package. Read about the data in the corresponding R help file. What are the variables included in this dataset? Take a look at the data using `head(iris)`.
2. What is the class of the iris dataset?
3. How many rows and columns does the `iris` dataset have? What R functions directly give the number of rows and columns?
4. Read the help page for $ (use `?"$"`). How can you use $ to subset the `iris` dataset? Use $ to compute the mean sepal length in the iris dataset.
5. What is the R function for standard deviation? Read the help page for this function. Compute the standard deviation of sepal width in the iris dataset.
6. What is the R function for obtaining quantiles of a vector? Compute the 20th and 80th percentiles of petal length.

Part 2. Google flu data

This data, introduced in lecture, estimates 2013 US flu activity using google searches. The data includes estimated daily flu activity for the whole United States, Georgia, Atlanta, and Health and Human Services Region 4 (which includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee). Google published their results in Nature (http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html). An article in Science also discusses problems with using this data to predict flu activity (http://www.sciencemag.org/content/343/6176/1203).

Data Source: Google Flu Trends (http://www.google.org/flutrends)

1. Download the dataset “googleflumissing.txt” from www.jennakrall.com/IntrotoRepi/googleflumissing.txt . Open the file in a text editor (e.g. TextEdit on a Mac, Notepad in Windows). What separates the entries in this dataset? See the help page in R for `read.table` and read this dataset into R.
2. Look at the data using `head()`

3. What symbol represents missing data in R? Look at the corresponding help page.

4. Find the mean of Georgia flu activity. What additional argument do you need?

5. Apply the `is.na` function to Georgia flu activity. Save the results of `is.na` to a new vector `na_georgia`. What class is `na_georgia`?

6. Take the mean of your vector `na_georgia`. What value does it give and what does this correspond to?

7. The `which` function in R gives the locations of elements within an R object that meet a certain criteria. Use this function to determine which elements are missing Georgia flu activity.

8. See the R help page for ! (`?"!"`). Use this operator to create a vector `notna_georgia` that indicates the values that are not missing Georgia flu activity.

9. Use your vector `notna_georgia` to create a vector of the Georgia flu activity for only those days that are not missing.

10. Using the principles in #9, find the mean flu activity in Georgia when the Atlanta flu activity is above its median.

11. What is the class of the Date variable in the flu dataset? What happens when you compute the mean Date?

12. The flu season is generally from October through April. Create a new variable indicating when it is flu season. Hint: To create a date, use `as.Date`. You may also need to look again at logical operators in R (`?"|"`).

13. The variable `flu_high` indicates when the number of flu searches in Georgia was above 3500. What class is this variable? What are its levels?

14. Change the labelling of `flu_high` using only the `factor` function to be “> 3500” for high flu search days and “<= 3500” for low flu search days.